TimeChat: A Time-sensitive Multimodal Large Language Model for Long Video Understanding

Year: 2023 | *Citations: 328* | *Authors: Shuhuai Ren, Linli Yao, Shicheng Li*

Abstract

This work proposes TimeChat, a time-sensitive multi-modal large language model specifically designed for long video understanding. Our model incorporates two key architectural contributions: (1) a timestamp-aware frame encoder that binds visual content with the timestamp of each frame, and (2) a sliding video Q-Former that produces a video token sequence of varying lengths to accommodate videos of various durations. Additionally, we construct an instruction-tuning dataset, encompassing 6 tasks and a total of 125K instances, to further enhance TimeChat's instruction-following performance. Experiment results across various video understanding tasks, such as dense captioning, temporal grounding, and highlight detection, demonstrate TimeChat's strong zero-shot temporal localization and reasoning capabilities. For example, it achieves +9.2 F1 score and +2.8 CIDEr on YouCook2, +5.8 HIT@1 on QVHighlights, and +27.5 R@1 (I oU=0.5) on Charades-STA, compared to state-of-the-art video large language models, holding the potential to serve as a versatile video assistant for long-form video comprehension tasks and satisfy realistic user requirements.11Our code and dataset are available at https://github.com/RenShuhuai-Andy/TimeChat.